RobustMark: Robust and Invisible Watermarking via Bit Distribution Balancing

Geunwoo Oh, Choongsang Cho, Guisik Kim
Intelligent Image Processing Research Center
Korea Electronics Technology Institute (KETI)
Seongnam, South Korea
{geunwoo0503, ideafisher, specialre}@keti.re.kr

***Abstract*—The rapid advance of ICT and generative AI has highlighted the need to protect the authenticity of digital content. Invisible watermarking offers an effective solution by imperceptibly concealing information in images. We present RobustMark, an invisible watermarking method designed for robust and imperceptible embedding and extraction of the watermark, regardless of the watermark bit distribution. To enhance robustness for non-uniform watermark inputs, we introduce a Bit Distribution Equalizer (BDE) module that transforms the distribution of the watermark bit sequence to be uniform by applying a bitwise XOR operation with an internally stored uniformly random bit sequence. Experiments with various watermark inputs under diverse image transformations demonstrate that RobustMark maintains high imperceptibility and achieves robust recovery performance even for non-uniform watermarks.**

***Index Terms*—Invisible watermarking, steganography, security, neural network.**

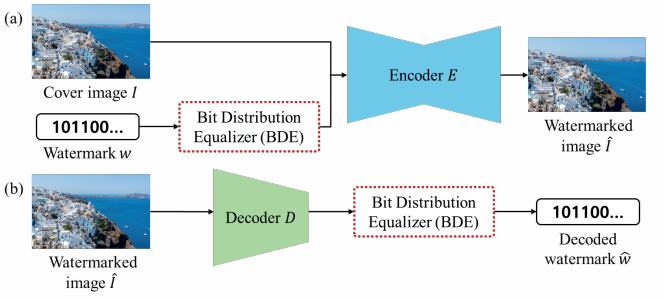

Fig. 1. Overview of proposed method: (a) Encoder embeds watermark, whose distribution has been equalized by the bit distribution equalizer (BDE), into the image. (b) Decoder extracts the watermark from the images and converts it using the BDE.

I. Introduction

The spread of ICT convergence and the rapid advancement of generative AI have underscored the critical importance of protecting the authenticity of digital content and preventing misinformation. Generative AI can produce highly realistic images, which may be misused for malicious purposes, such as creating deepfakes, interfering with elections, and undermining social infrastructure. Invisible watermarking is an effective solution that imperceptibly conceals information in images to enable reliable verification while preserving image quality.

Classical watermarking techniques, which embed information in the least significant bit (LSB) or the frequency domain, are vulnerable to a small modification to the encoded image. Recently, deep learning-based approaches have improved robustness to modifications on the encoded image while preserving the image quality. These advances support a high-resolution encoded image and larger watermark payloads [1].

In practice, users may embed not only watermarks with uniformly distributed bits, e.g., UUID, but also arbitrary data, such as alphabets or numbers, by converting them into bit sequences. These converted bit sequences may exhibit imbalanced distributions of '0' and '1' bits. Previous deep learning-based watermarking methods typically assume uniformly random bit distributions, leading to degraded performance when faced with such imbalanced conditions.

To address this problem, we propose RobustMark, which balances the distribution of watermark bits and enables reliable extraction even under non-uniform bit distributions. To this end, we introduce the Bit Distribution Equalizer (BDE) module to ensure robust recovery of the watermark.

II. Methodology

A. Watermark Encoder and Decoder

The watermark encoder network E takes the watermark bit sequence $w \in \{0,1\}^l$ (where l indicates the length of the bit sequence) and the cover image I as input, and outputs the watermarked image $\hat{I}$. Following InvisMark [1], we preprocess the watermark bit sequence into a tensor, concatenate it with the resized cover image, and postprocess the generated watermark residual by upscaling to match the cover image's resolution and then adding to it. In our encoder network, we exploit depthwise convolution to improve parameter efficiency and employ pixel shuffle and unshuffle operations for efficient and effective spatial resolution scaling. The decoder network D, based on ConvNeXt, recovers the watermark bit sequence $\hat{w} = D(\hat{I})$ from the watermarked image $\hat{I}$.

B. Bit Distribution Equalizer (BDE)

The Bit Distribution Equalizer (BDE) transforms the distribution of the watermark bit sequence to be uniform by performing a bitwise XOR operation between the user-provided watermark bit sequence and an internally stored uniformly random one. Since the XOR operation is applied with a uniformly random sequence, the output of BDE also exhibits a uniform distribution. To facilitate more effective embedding by the encoder, BDE converts even a highly imbalanced

TABLE I

WATERMARK BIT ACCURACY IN VARIOUS IMAGE TRANSFORMATIONS ON THE DIV2K DATASET. EACH COLUMN CORRESPONDS TO A DIFFERENT INPUT TYPE: RANDOM UUID, RANDOM ALPHABET SEQUENCE, RANDOM NUMBER SEQUENCE, AND A USER-PROVIDED INPUT ("ICTC 2025" AS AN EXAMPLE).

	Random UUID		Random Alphabet			Random Number			ICTC 2025		
	InvisMark	RobustMark	InvisMark	RobustMark	RobustMark w/ BDE	InvisMark	RobustMark	RobustMark w/ BDE	InvisMark	RobustMark	RobustMark w/ BDE
Clean	100.00	100.00	96.14	96.45	99.43	90.07	92.41	97.88	86.56	90.40	97.96
Brightness	99.34	99.59	94.48	95.53	99.03	87.10	89.94	96.90	82.73	86.45	97.07
Contrast	99.94	99.94	95.84	96.28	99.37	89.28	91.87	97.71	85.62	89.34	97.83
Saturation	100.00	99.99	96.08	96.43	99.42	89.97	92.28	97.85	86.49	90.17	97.94
Jiggle	99.97	99.99	95.98	96.39	99.42	89.57	92.06	97.80	86.04	89.70	97.88
Jpeg	99.57	99.39	94.02	94.65	98.71	85.58	87.75	96.01	78.77	82.38	96.19
Posterize	99.99	99.99	96.04	96.40	99.43	89.79	92.18	97.82	86.17	89.83	97.93
RGB Shift	100.00	99.99	96.08	96.42	99.42	89.92	92.30	97.84	86.37	90.16	97.94
Gaussian Noise	100.00	100.00	95.87	96.22	99.43	88.73	91.04	97.60	84.39	87.36	97.81
Gaussian Blur	100.00	100.00	96.13	96.46	99.43	90.08	92.45	97.88	86.58	90.39	97.96
Box Blur	100.00	100.00	96.12	96.46	99.43	90.10	92.48	97.87	86.60	90.35	97.96
Flip	99.99	99.98	95.91	96.29	99.41	88.52	91.33	97.78	84.90	87.88	97.79
Perspective	99.91	99.97	95.08	96.12	99.40	86.87	90.63	97.59	83.59	87.37	97.83
Random Erasing	99.93	99.89	95.43	95.98	99.34	88.15	90.66	97.42	84.92	88.17	97.50
Random Resized Crop	98.02	96.65	92.92	93.48	96.12	85.33	87.37	94.47	82.01	84.34	94.75

TABLE II

COMPARISON FOR WATERMARK IMPERCEPTIBILITY BASED ON PSNR AND SSIM ON THE DIV2K DATASET.

	InvisMark	RobustMark
PSNR	49.28	52.60
SSIM	0.9980	0.9989

input bit sequence into a uniform one. During extraction, the BDE enables the decoder's output to be converted back to the original watermark bit sequence by applying the XOR operation with the same random sequence.

III. EXPERIMENTAL RESULTS

We train our model on the DALL·E 3 dataset [2], an open-source collection of generative AI images, and validate it using the validation set from the CLIC dataset [3]. We embed randomly generated 100-bit watermark and follow the same training settings of InvisMark. Finally, we evaluate the performance on the DIV2K [4] dataset, consisting of 2K resolution images, using all 900 images for evaluation. To measure the imperceptibility of the encoded images, we employ peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). To assess watermark recovery, we report bit accuracy between the decoded and the input watermark.

We report imperceptibility of watermarks using image quality scores, measured on the DIV2K dataset in Tab II. RobustMark outperforms the state-of-the-art InvisMark in terms of image quality, indicating that watermarks embedded by RobustMark are less perceptible while maintaining high visual fidelity. Tab. I presents the decoded watermark bit accuracy under various image transformations on the DIV2K dataset, using the same transformation setting of InvisMark. For random universally unique identifier (UUID) inputs, which exhibit uniform bit distribution, RobustMark performs comparably to or better than InvisMark. In contrast, random alphabet and number sequences exhibit non-uniform bit distributions, as the underlying ASCII codes of letters and digits are inherently imbalanced in their binary representation due to shared fixed high-order bits. In these cases, RobustMark achieves higher bit accuracy than InvisMark. In particular, RobustMark with BDE further enhances robustness for non-uniform inputs. These results indicate that balancing the bit distribution through BDE significantly improves the robustness of watermark recovery.

IV. CONCLUSION

We propose RobustMark, a robust invisible watermarking method that enables embedding and extracting the watermark reliably, even when the watermark bit distribution is imbalanced. We introduce the Bit Distribution Equalizer (BDE) to balance the watermark bit distribution. By incorporating BDE, our method transforms any input watermark bit sequence into a uniform distribution, enabling stable and effective embedding even for highly imbalanced inputs. Experiments with various input types and image transformations demonstrate that our method enables stable and robust watermark embedding and extraction. As a result, RobustMark enables user-provided watermarks to be embedded and extracted with enhanced robustness, indicating that it is well-suited for practical invisible watermarking for digital content protection.

ACKNOWLEDGMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2024-00456709), and the Ministry of Culture, Sports and Tourism (MCST) (RS-2024-00441174).

REFERENCES

[1] R. Xu, M. Hu, D. Lei, Y. Li, D. Lowe, A. Gorevski, M. Wang, E. Ching, and A. Deng, "Invismark: Invisible and robust watermarking for ai-generated image provenance," in *IEEE Conf. Winter Conf. Appl. Comput. Vis.*, pp. 909–918, 2025.

[2] "Dall·e 3 dataset." https://huggingface.co/datasets/OpenDatasets/dalle-3-dataset.

[3] G. Toderici, W. Shi, R. Timofte, L. Theis, J. Balle, E. Agustsson, N. Johnston, and F. Mentzer, "Workshop and challenge on learned image compression (clic2020)," 2020. http://www.compression.cc.

[4] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, July 2017.